Measuring the Effects of Starting Pitching

Lee Przybylski

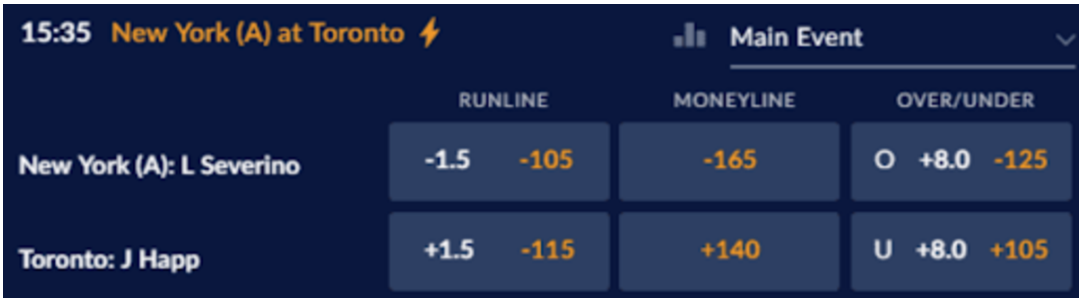
November 17, 2021

Abstract

Betting on baseball is challenging. One feature that makes the sport different is that moneylines usually list probable starting pitchers. To take advantage of this, we develop a generalized linear mixed effects model using retrosheet data from several seasons. The model includes effects for teams, starting pitchers, and venue. Being able to assess a pitcher's performance independent of his team is also challenging. By estimating effects for each starting pitcher, fitting the model provides another way measure a starting pitcher's effectiveness. We also provide some background on pitching metrics that have been used in the past, such as ERA, FIP, and oppent WOBA, and compare these metrics to our estimated pitcher effects.

1 Introduction

Most of our data will be taken from sportsbookreviewsonline.com



2 Model Selection

2.1 Overdispersion

When we use regression to fit a generalized linear model (GLM) with an exponential family of distributions, it is usually the case that the variance of our distribution is a function of the mean of the distribution. Recall that a Poisson distribution with mean $\lambda > 0$, denoted poiss(λ), has pmf

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \mathbb{N}_0.$$

Through a link function, we can use ordinary least squares to find an estimate for the mean based on the covariates, but often variance implied by this estimated mean is too small to explain the variation observed in the data. When this happens, we say that the model has *overdispersion*. This must be accounted for in order to perform meaningful inference. There are 2 common ways to deal with overdispersion.

- We can add a dispersion parameter to the model which is estimated using a quasi-likelihood approach.
- We can fit a generalized linear mixed effects model (GLMEM) where we assume the mean is also affected by a centered normal random variable.

Here we will elect to follow the second approach.

2.2 Predictive Value of the Model

Let y_{ijklm} be the number of runs scored by team i against team j at venue k facing starting pitcher l during the mth game of the season. The model we propose assumes that $y_{ijklm} \sim \text{Poisson}(\lambda_{ijklm})$ where

$$\log(\lambda_{ijklm}) = \mu + \chi \mathbf{1}_{im} + b_i + f_j + v_k + p_l + g_m + e_{im}, \tag{1}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2), \;\; g_m \overset{\text{iid}}{\sim} N(0, \sigma_g^2), e_{im} \overset{\text{iid}}{\sim} N(0, \sigma_e^2).$$

$$\mathbf{1}_{im} = \begin{cases} 1 & \text{if team } i \text{ is home during game } m \\ 0 & \text{otherwise} \end{cases}$$

3 Pitcher Effects

3.1 Noteworthy Pitchers

3.2 Comparison with Other Metrics